Kuin, NIKI., Masthoff, E., Munafo, M., Nunnink, V., & Penton-Voak, I. (Accepted/In press). Changing perception: A randomized controlled trial of emotion recognition training to reduce anger and aggression in violent offenders. *Psychology of Violence*.

Changing perception:

A randomized controlled trial of facial emotion recognition training in order to reduce anger and aggression in violent offenders

Abstract

Objective: To determine whether emotion recognition training, which previously proved to be effective in adolescents, also reduces anger and aggression in adult violent offenders.

Method: Detained male adults were randomized to complete either a 1-week computer training designed to promote the perception of happiness over anger in ambiguous facial expressions (n = 46), or a sham training control procedure (n = 44). Outcome measures were collected immediately after training and at 6-week follow-up, and included the number of faces that were rated as happy rather than angry, self-reported and observed measures of hostility, aggression and prosocial behaviour. The linear regression analyses were statistically corrected for age and presence of (mild) intellectual disability.

Results: The training procedure was highly effective in promoting the perception of happiness over anger in the training group as compared to the controls, independent of age or intelligence (95% CI -4.6 to -2.8, $p < 0.001$). These training effects remained at six weeks post training (95% CI -3.4 to -1.8, $p < 0.001$). There was no clear change in measures of aggression and hostility, or prosocial behaviour.

Conclusions: In contrast to two prior studies with adolescent samples, the present study showed no meaningful impact of the training procedure on aggression in adult offenders, even though the training was effective in altering emotion perception. This may be due to low statistical power, or a lack of generalization of perception of happiness to faces in daily life encounters, or because emotion recognition bias is not causally related to aggression.

Keywords: aggression treatment, emotion recognition, hostile attribution bias, randomized controlled trial, offenders

Public health significance statement

Since aggression and violence are globally recognized problems for personal wellbeing and society, it is important to develop more effective interventions to reduce aggressive behaviour. In the present randomized controlled study a new treatment approach was investigated, in which offenders learned to interpret facial expressions of others as happy rather than angry. Although this computer training was successful in changing emotion recognition, this was not accompanied by a decline in measures of aggression.

Introduction

Aggression and violence have a negative impact on both society and personal wellbeing, often leading to major negative consequences for both victims and aggressors (Lee, 2016). Despite the existence of multiple treatment programs to reduce and prevent violent behaviour, the need for more effective and targeted interventions remains (Lee et al., 2016; Mikton, Butchart, Dahlberg, & Krug, 2016). To obtain more insight into the factors involved in initiation and perpetuation of aggression and violence, and to find ways to decrease aggressive behaviour, there is ongoing interest in understanding underlying neurocognitive and neurobiological mechanisms (Angus, Schutter, Terburg, Van Honk, & Harmon-Jones, 2016; Dean, 2014). This includes, for example, aspects of social cognition including facial emotion perception (Marsh & Blair, 2008). A robust association appears to exist between antisocial behaviour and deficits in recognizing specific facial emotional expressions of which the inability to correctly perceive fearful expressions is the most pronounced (Marsh & Blair, 2008). In addition to these general findings, other studies have

focused on more specific biases in facial emotion perception in relation to aggression and violence, such as the 'hostile interpretation bias'. This bias reflects the tendency for aggression-prone individuals to perceive emotional facial expressions as more angry or hostile than non-aggressive individuals. Such a tendency appears to exist not only in the perception of ambiguous faces (Schönenberg & Jusyte, 2014), but is shown to be generalized to more clear, less ambiguous emotional intensities as well (Smeijers, Rinck, Bulten, van den Heuvel, & Verkes, 2017). The presence of this hostile interpretation bias may have important clinical relevance for the maintenance of aggressive behaviour, because it increases the chance of creating a spiral of hostility through a self-reinforcing mechanism. When someone perceives another as hostile, they might initiate social interactions with a more hostile stance, leading to a more hostile response in return. The opposite might be true as well: a tendency to perceive others as more happy and friendly might elicit more pro-social behaviour and therefore lead to more self-reinforcing positive interactions. In line with this reasoning, the present study focusses on the reversibility of such a bias in treatment and its subsequent effect on aggressive characteristics. Two promising studies in aggression-prone youths have already found evidence for the existence - and possibly even reversibility - of a causal relationship between hostile interpretation of facial expressions and aggressive tendencies (Penton-Voak et al., 2013; Stoddard et al., 2016). These studies showed that it is possible to reduce the tendency for hostile interpretation of ambiguous facial expressions through training. In one study this was found to result in a reduction in self-reported and observed anger, irritability and aggression. This positive training effect increased further after the training was ended, presumably because of a self-enhancing, positive feedback mechanism, which was elicited as a result of the training (Penton-Voak et al., 2013). A similar effect has been found in a normal student population after training (AlMoghrabi, Huijding, & Franken, 2018).

Commented [N1]: Evt referenties downsizen, of gehele paragraaf vanaf 'a similar effect weglaten'

These studies provide a basis for further research into the effectiveness of training programs to reduce hostile interpretation biases in order to reduce aggressive behaviour. In these studies the high risk groups studied comprised adolescents (Penton-Voak et al., 2013; Stoddard et al., 2016), but age might be an important factor to consider in generalization of findings from these training studies as hostile interpretation bias may decline with age (Kuin, Masthoff, Munafò, & Penton-Voak, 2017). This might make training more relevant and effective in adolescents than in adults at risk for aggressive behaviour. Furthermore, because increased positive interactions may enhance any effects of training, there is a need to study prosocial behaviour in addition to measures of aggression. Relatedly, it has been suggested that measuring prosocial behaviour may be important for detecting the effectiveness of interventions in clinical forensic samples, rather than measures of aggression, because the prevalence of aggressive behaviour is usually low during imprisonment to begin with (Hornsveld, Nijman, Hollin, & Kraaimaat, 2007). Finally, more research is needed to determine whether this type of training is effective for specific highly prevalent forensic subgroups, such as people with mild intellectual disability (Fazel, Xenitidis, & Powell, 2008). Offenders with mild intellectual disability may be less responsive to traditional verbal interventions, and may benefit particularly from implicit learning strategies (Lisle, 2007; Marotta, 2017).

The present randomized controlled study was designed to gain more insight into these issues. Our primary aim was to determine whether an emotion recognition training procedure designed to promote the perception of happiness over anger in ambiguous emotional expressions results in: 1) a reduction of hostile interpretation bias among adult male detained offenders, 2) a decline in self-perceived and observed anger and aggressive behaviour, and 3) an increase in prosocial behaviour. A secondary objective was to gain insight into the relative effectiveness of the intervention for offenders with mild intellectual disability as opposed to

offenders with average intelligence. Because of the implicit, non-verbal nature of the training we hypothesized that the training would be equally successful for participants with estimated low as well as normal intelligence.

Materials and Methods

Setting and Participants

The study was conducted in the Penitentiary Institution in Vught, one of the larger correctional facilities in the Netherlands, where both regular prison, specialized forensic psychiatric, and high security wards are located. All participants were adult male offenders, who were detained for a variety of offenses, ranging from minor crimes to severe violent crimes, including sex crimes. Since a previous study showed no clear evidence of an association between offense type and performance on the emotion perception task (Kuin et al., 2017), all types of offenders were included. However, the vast majority of the study population (97%) had committed a violent crime at least once in their lifetime. During their participation in the study they resided in either regular wards (24% of the experimental group and 34% of the control group), the psychiatric treatment centre of the prison (53% of the experimental group and 29% of the control group) or a specialized section for repeated offenders (22% and 37% of the experimental and control group respectively). Participants were excluded from the study if they were diagnosed with an autism spectrum disorder or with an active episode of a severe psychiatric disorder (psychotic, bipolar or major depressive disorder) in the three months prior to participation. Additionally, participants were excluded if staff members expressed concern about safety issues (for example, high risk for aggression or major disruptive behaviour during the training). In order to be able to complete the required forms and undergo testing, only participants were included who were well acquainted with the Dutch language (though not necessarily native speaking) and who had remaining sentences of at least eleven weeks (in order to be able to complete the study). Nevertheless, some of the

participants dropped out due to unforeseen release or transfer to other prisons. Figure 1 shows the number of participants and drop-out rates in both groups throughout the study. Characteristics of participants, who were included in the statistical analyses, are presented in Table 1.

The study was approved by the scientific department of the Dutch Ministry of Justice and Security with respect to procedural and ethical aspects. All participants signed for informed consent after receiving both verbal and written information about the study.

Procedure

We used a double blind two-arm randomised placebo-control design. All participants were tested over eleven weeks, across three phases that were completed in a fixed order. The first phase was a pre-testing period of 4 weeks, the second phase a training week when the actual intervention was completed, and the third phase a six-week follow-up period. More details on specific activities in each phase are provided below. The study was conducted over a period of approximately 1,5 years, across a total of eight partially overlapping waves. In each wave, approximately ten to twelve participants were included..

Participation in the study was voluntary. Research assistants recruited detained participants in the institution through posters and information letters, explaining the aims of the study, as well as through personal contact. The information letter explicitly stated that the study was conducted in order to determine the effectivity of a new training program to reduce aggression and that this involved a training in the perception of facial emotional expressions. It was also explained to all participants in advance that there were two conditions with one training condition and one placebo condition and that participants would not be informed about which condition they were assigned to. Staff members and psychologists were consulted to assess whether inclusion criteria were met. Suitable candidates were then approached individually to inform them about the study and invite them to participate. Further

written information was subsequently provided, and a new appointment was planned to give participants deliberation time. In that second appointment written consent was signed followed by a short intake interview to assess basic participant characteristics (e.g., age and education). Participants were randomly allocated to either the experimental condition or the control condition by means of a randomization tool (from http://www.randomization.com/), which uses randomly permuted blocks and is based on the modified pseudo-random number generator from Wichmann and Hill (1982) (McLeod, 1985). Neither participants nor trainers knew the allocated condition, although they were both explicitly informed about the fact that there was a placebo condition. Participants in both conditions completed the training phase in mixed groups of approximately six participants, so conditions were identical for participants from both groups. Based on individual computer codes, entered by the trainer, the computer started either the experimental training or the control training procedure which were visually indistinguishable. Trainers were not informed about which code represented which condition, to assure the double-blind design. During five consecutive days (Monday to Friday) the training sessions took place. Each session took approximately 30 to 45 minutes to complete.

No incentives were provided for participation in the study. However, to enhance motivation during training sessions, free soda and cookies were provided to the participants.

The trainers were all master students clinical (neuro)psychology. They were not only present as trainers during the training, but were also responsible for recruitment of participants and they visited participants weekly to distribute and collect questionnaires during the complete study trajectory.

The trial protocol was not pre-registered.

Instruments

The experimental intervention paradigm:

The intervention, an emotion recognition training procedure, was a computer-based task designed to modify the perception of ambiguous facial expressions of emotion, originally developed by Penton-Voak et al. (2013) (see Figure 2). Prototypical happy and angry composite images were derived from 20 individual male faces showing a happy facial expression and the same 20 individuals showing an angry expression. The original images came from the Karolinska Directed Emotional Faces (Lundqvist, Flykt & Öhman, 1998). These prototypical images were used as endpoints to generate a linear morph sequence that consists of fifteen images that change incrementally from unambiguously happy to unambiguously angry, with emotionally ambiguous images in the middle. During each training session participants were instructed to rate these images as either happy or angry, in a two-alternative forced-choice procedure administered by a computerised test in E-Prime 2.0.

First a fixation cross appeared (for 1500-2500 milliseconds, randomly jittered), followed by a short presentation of one of the faces on the happy-angry continuum (for 150 milliseconds), and then by a mask of visual noise (150 milliseconds), at which point participants rated the face as either happy (by pressing 'C') or angry (by pressing 'M'). The mask was presented to disrupt processing of visual afterimages, and so judgements relied on processing of the brief presentation of the emotional expression. At the beginning of each training session a baseline was calculated: a simple estimate of each participant's balance point between happy and angry responses. This 'threshold score' was derived by counting the proportion of 'happy' responses as a proportion of the total number of trials (Penton-Voak et al, 2013). In the emotion perception task a lower threshold score reflect a tendency to rate the faces as angry (i.e. the participant considers a smaller number of faces as happy), while higher scores indicate that a larger proportion of the continuum is perceived as happy. Although construct validity of the task was not explicitly assessed, it has been shown to differentiate between youths with disruptive mood dysregulation disorders and healthy controls (Stoddard et al.,

2016). Baseline blocks consisted of 45 trials, with each face from the 15-face continuum presented three times.

After this baseline assessment, the training started. This training was designed to encourage positive interpretations of ambiguous facial expressions. Each trial in the training phase was identical to baseline trials with respect to the inter-trial interval and stimulus presentation, but after participants responded feedback was given. In the control condition, feedback was directly based on the participant's baseline balance point. So, responses were classified as "correct" when the participant identified images below the original balance-point image as happy, and faces above that image as angry, and otherwise were classified as "incorrect" (i.e. "incorrect" feedback was given when the response given was inconsistent with the threshold score calculated from that participant's baseline performance). Feedback was a message saying, "Correct/Incorrect! That face was happy/angry" combined with a non-verbal visual cue (a green checkmark for correct responses and a red cross for incorrect responses). In the experimental condition, feedback was also based on the participant's baseline balance point, but the "correct" classification was shifted two morph steps toward the angry end of the continuum, so that the two images nearest the balance point that the participant would have classified as angry at baseline were considered happy for purposes of feedback (Penton-Voak et al, 2013, see Figure 2). Each block of training consisted of 31 trials, with the four most unambiguous images presented once (images 1,2,14,15), the six intermediately ambiguous images presented twice (images 3,4,5,11,12,13) and the most ambiguous images presented three times (images 6,7,8,9,10). Six blocks of training were presented in each session. Following training, a 'test' block (identical to the baseline block) was administered to assess whether training had changed responses made to faces.

The initial baseline measure on the first training day (e.g. before any training took place) was used for analysis, as well as the final 'test' measure at the fifth (final) training day.

Thereupon, the baseline measure was repeated at three and six weeks follow-up to determine the resilience of training effects on emotion perception (no training sessions were administered at these time points).

Outcome Measures:

Self-report questionnaires. Self-report measures were used to assess anger, hostility and aggression. During the complete eleven-week period ranging from four weeks before pre-training to six weeks post-training, participants rated their own aggression weekly on a short 12-item form with a 5-point Likert scale ranging from '0: not present' to '4: (almost) always present', further referred to as *'Self-Report'* (primary outcome). Items reflected concrete behaviour, reflecting both verbal and physical aggression, as well as hostile perceptions or feelings of anger or irritability. Scores were summed to generate a total score. This self-report questionnaire was largely based on the Social Dysfunction and Aggression Scale (SDAS-11; Wistedt et al., 1990). It has been shown to have good applicability and convergent validity and moderate inter-rater reliability in Dutch forensic settings (Bousardt, Hoogendoorn, Noorthoorn, Hummelen, & Nijman, 2016; Kobes, Nijman, & Bulten, 2012). Most items were reformulated for better understanding for people with low intellectual capacities, two items on suicidality and self-harm were removed and three items were added on feelings of provocation and inhibition of aggression. The internal consistency of this instrument was good in the present sample (Cronbach's α = .839).

The *Novaco Anger Scale – Provocation Inventory* (NAS-PI; Novaco, 1994) is a questionnaire, providing two scores estimating feelings of anger (first 48 items, NAS) and sensitivity to provocation (last 25 items, PI). The Dutch translation has good internal consistency, test-retest reliability and validity (Hornsveld, Muris, & Kraaimaat, 2011). The NAS-PI was assessed three times: directly pre- and post-training and at six weeks follow-up.

Behavioural observations: The *Observation Scale for Aggressive Behavior* (OSAB; Hornsveld, Nijman, Hollin, & Kraaimaat, 2007) was applied for structured observations of aggressive and prosocial behaviour during eleven weeks (four weeks pre-training, during training and six weeks post-training). Each week, one staff member rated the frequency of specific behaviour, observed during the preceding week on 40 items with a four-point scale (response options vary from 'no' to 'frequently'). Of the six available outcome scales, only the scales 'Irritation/anger', 'Aggressive Behavior' and 'Prosocial Behavior' were applied in the present study. The Prosocial Behavior scale was specifically included in this study, because it has been suggested that in closed, highly secured and structured forensic settings (such as in the present study) it may be easier to detect behavioural progress through increased ratings of prosocial behaviour instead of through decline in aggressive behaviour (Hornsveld et al., 2007). The OSAB's good internal consistency and test-retest reliability have been confirmed in a study with Dutch violent forensic psychiatric patients (Hornsveld et al., 2007).

Other measures. The Screener for Intelligence and Intellectual Disability (SCIL; Kaal, Nijman, & Moonen, 2012) is a screening tool to assess intelligence, which was developed in the Netherlands for intelligence-screening with adolescents and adults in forensic care settings. Assessment of the SCIL starts with a short interview asking about, for example, educational level and prior healthcare referrals for people with intellectual disabilities. This is followed by short assignments, such as simple arithmetic tasks, reading and writing tasks, clock drawing, etc. It takes approximately fifteen minutes to complete and has good psychometric properties in both Dutch adult and juvenile populations (Nijman, Kaal, Scheppingen, & Moonen, 2018) and is well applicable in the Dutch prison system (H. L. Kaal, Nijman, & Moonen, 2015). The cut-off for the SCIL (score <19.5) provides a rough estimation whether a respondent has an IQ below 85.

Judicial records were screened to obtain insight in conviction histories. These data were only collected to provide descriptive information about the participants, but were not included in statistical analyses, because it was not yet possible to assess if training had effect on actual criminal violent behaviour.

Statistical procedure

For practical reasons the number of participants was fixed at approximately 80 (40 participants per group). We calculated that this would allow us to detect an effect size *(d*) of 0.56 with 80% power at a 5% alpha level, equivalent to an approximately 3-point difference on our primary outcome (the self-reported aggression score), assuming a *sd* of 5.26.

The threshold scores on the emotion perception task at baseline were screened for outliers that could point to an invalid, random response style. Data of participants with extreme high (≥+2,5 *sd*) or low baseline threshold scores (≤-2,5 *sd*) were excluded from further analysis.

No imputations were made because missing data were spread completely at random across the sample. Since data on self-reported and observed aggressive behaviour variables were not normally distributed, these were all transformed using a natural log before the linear regression analyses. There were no problems with multicollinearity in the data. Linear regression analyses were used to assess the relationship between training condition and outcome measures on emotion perception and aspects of aggression. These regression analyses were minimally adjusted at first (i.e., only adjusted for baseline), and subsequently fully adjusted (for baseline, age, intelligence, and for a potential moderator effect of age x condition). Included outcome variables of aggression were staff-rated anger/irritation, aggression and prosocial behaviour (all assessed by the OSAB), self-reported aggression, and self-reported, hostility and anger (NAS-PI). Self-reported aggression and staff-observed anger, aggression and prosocial behaviour were all assessed on a weekly basis, starting four weeks before training

up to six weeks post training. Mean scores were calculated for the pre-training period for each scale (the baseline score). A minimum of two measures in this four-week period was necessary to be included for further analysis. In a similar manner, a single post-training score was derived for each scale calculating mean scores of measures during the six-week follow-up period. A minimum of three measures was required to be included for further analysis.

To establish if training effects were similar for people with or without estimated mild intellectual disability, three ANOVA's were conducted within the experimental group, comparing two subgroups of participants with or without estimated mild intellectual disability on three measures of the threshold (baseline, directly after training and at six weeks follow-up.

Results

Training effect on emotion recognition

Threshold scores on the emotion perception task were assessed at baseline, directly post-training and at six-weeks follow-up. Means scores and standard deviations for both groups are displayed in figure 3. Three participants in the control group and one in the experimental group were excluded from the analyses due to probable invalid response styles. Exclusion of those participants did not lead to meaningful differences in further statistical outcome.

In the fully adjusted linear regression analysis, participants in the intervention group had shifted their threshold by 3.7 frames on average (95% CI -4.6 to -2.8, $p < 0.001$) relative to those in the control group at the end of the 1-week period. After 6-weeks the mean difference was 2.6 frames (95% CI -3.4 to -1.8, $p < 0.001$). Age and having an estimated intellectual disability or not (based on SCIL-scores) did not contribute significantly in the prediction of the threshold scores after training.

Training effect in participants with estimated low IQ

Based on the SCIL 22 out of 45 participants in the experimental group were estimated to have IQ-scores below 85. Those 22 participants were compared to the other 23 participants in the experimental condition on their mean threshold scores across the three measuring moments (see Figure 4). There was a significant difference between threshold scores in those groups at baseline, $F(1, 43) = 4.16$, $p = .048$, and at six-weeks follow-up, $F(1, 38) = 4.56$, $p = .039$. Directly after the training these subgroups had equal threshold scores, $F(1, 43) = .48$, $p = .491$.

Training effect on measures of anger, hostility, aggression and prosocial behaviour

Mean scores and standard deviations for each aggression variable at baseline are displayed in Table 1. Table 2 displays the differences on measures of aggression after training and the main regression coefficients, both minimally and fully adjusted.

There was no clear evidence of an effect of the intervention on any of our measures of aggression, hostility or prosocial behaviour. Estimated mild intellectual disability (based on SCIL measures) and age were not meaningful predictors of the outcome, and there was no clear evidence of an age x condition interaction effect in any analysis.

Discussion

The primary aim in the present study was to determine whether an emotion recognition training procedure designed to promote the perception of happiness over anger in ambiguous emotional expressions results in: 1) a reduction of hostile interpretation bias among adult male detained offenders, 2) a subsequent decline in self-perceived and observed anger and aggressive behaviour, and 3) an increase in prosocial behaviour. Our results indicate that the training was very successful in shifting ratings of ambiguous faces from

angry to happy, but that this was not accompanied by a meaningful change in self-perceived and observed anger, aggression or prosocial behaviour.

On average participants in the intervention group rated more faces as happy rather than angry after the training compared with those in the control group, and this effect remained almost the same after a period of six weeks. This is in line with the training effect found using the same procedure in other studies with adolescents (Penton-Voak et al., 2013; Stoddard et al., 2016), although follow-up in these studies was limited to two instead of six weeks. It is promising that these effects on the rating of emotional expressions appear to remain stable over a longer period of time, which suggests that such a computer training may be an effective means to reduce a hostile interpretation bias in adult as well as adolescent offenders.

A secondary objective in this study was to gain insight into the relative effectiveness of the intervention for offenders with mild intellectual disability as opposed to offenders with average intelligence. Our data showed that participants with estimated mild intellectual disability profit just as much from this computer training as do participants with higher cognitive ability, although the decline of the training effect after six week follow up is slightly stronger in the first group. This is an important finding, since the prevalence of mild intellectual disability in prisons is considerable (Fazel et al., 2008), and this group may be less responsive to traditional verbal psychotherapeutic interventions than people with higher intelligence levels (Cooney, Tunney, & O'Reilly, 2017; McGillivray, Gaskin, Newton, & Richardson, 2016; McNair, Woodrow, & Hare, 2017). Furthermore, the present study also provided weak evidence that participants with lower intellectual ability have a slightly more pronounced hostile interpretation bias at baseline than participants with (above) average intelligence, which emphasizes the vulnerability and need for treatment of this group.

Although these results with regard to the training effect on emotion perception are positive, the clinical relevance of such training in this context ultimately depends on its efficacy in reducing anger, aggression and violence. Our results indicate that participants in the experimental group showed no meaningful concurrent decline in anger or aggression as opposed to the controls as a result of the training, and the training did not contribute in the prediction of post-training measures of aggression. These findings are in contrast to previous intervention studies (AlMoghrabi et al., 2018; Penton-Voak et al., 2013; Stoddard et al., 2016), where a reduction in both self-reported and observed aggressive behaviour was clearly apparent in the intervention group. Two of those previous studies were conducted in youth with either aggression difficulties (Penton-Voak et al., 2013) or disruptive mood dysregulation disorder (Stoddard et al., 2016). A potential explanation for these differences could lie in age differences between the study groups, which could influence susceptibility to the training such that younger participants benefit to a greater degree than older participants. Moreover, age differences may not only be relevant with respect to aggression, but also in relation to hostile perception of emotions. In fact, in a previous explorative study with the same emotion perception task, was found that the tendency for hostile interpretation of facial expressions declined with age (Kuin et al., 2017), which may imply that the bias could have been less pronounced in the present population than in that of the previous two studies to begin with (this is further elaborated on in the limitations section below).

Another potential difference between the present study and the two previous ones is the setting where the participants resided. The participants in the two previous studies were not incarcerated or hospitalized during the study, in contrast to the detained males that took part in the present study. A prison setting is highly structured and restrictive (Ricciardelli & Memarpour, 2016). Interactions with other inmates and staff members are usually rather predictable and straightforward, which therefore may reduce the expression and subsequent

positive reinforcement of spontaneous social behaviour. In addition, because of the structured, secured and predictable prison-environment, base rates of aggression may be low, as was the case in this study as well. This makes measuring a potential decline in aggression challenging as result of a floor effect (Hornsveld et al., 2007). Precisely for this reason the present study incorporated measures of self-reported feelings of hostility and observational measures of prosocial behaviour, but, here also, no meaningful training effect in the expected direction was observed.

So, why is it that we failed to induce a decline in anger and aggression or an incline in prosocial behaviour, even when this training did appear effective in promoting the perception of happiness over anger? One potential explanation lies in the basic assumption of the existence of a (causal) relationship between the hostile interpretation of faces on the one hand and aspects of aggression on the other hand, which may not be as clear-cut to begin with. Evidence for the absence of such a relationship can be found in two prior studies that showed no significant correlations between aspects of aggression and a hostile interpretation bias (Kuin et al., 2017; Schwenk et al., 2014). In one of these studies the same emotion recognition paradigm was applied as in the present study (Kuin et al., 2017). And even if there is a relation between aggression and hostile interpretation of faces, such emotional perception problems might only be a symptom of aggression instead of a cause. In that line of reasoning an immediate decline in other aggressive symptoms after targeting hostile interpretations in the perception of facial expressions would not be obvious. Furthermore, it should be stressed that aggression can be caused by many individual and environmental factors, beside hostile interpretations (DeWall & Anderson, 2011). The assumption that a single focus in treatment on emotion perception, without regard to these other factors, could be enough to lead to a meaningful change on a behavioural level might be an unjustified oversimplification of the aggression-concept.

Study limitations

Even though the present study samples were relatively large for clinical samples in forensic populations, it appears as though the study was still underpowered. Although we saw a trend pointing to a decline of average scores on self-reported anger and aggression in the experimental group, while those of the control group increased, variances in these groups were too high to draw any plausible conclusions out of these findings. This trend could point to a true effect, but one that is too small to detect with the current sample sizes. In fact, the difference between both groups on this self-report measure was indeed slightly lower than the minimal difference we needed to find as indicated by our power calculation.

A second point to consider with regard to limitations of the present study is the possibility of a selection bias. For example, it could be that only individuals with high pro-social traits or lack of aggression problems volunteered for this study. The fact that 97% of all participants had been convicted for at least one violent crime in their lifetime and that participants in both groups had been convicted for an average of approximately eight violent crimes, rules out te possibility that only those offenders applied who had no problems with aggression. Furthermore, it was not tested to normative data whether the participants in the present study actually had a hostile interpretation bias to begin with. However, a previous study with the same training procedure as applied in the present study showed that this training was effective to reduce anger and aspects of aggression in a healthy young adult sample without objective perception biases or problems with aggression (Penton-Voak et al., 2013), which leads to the conclusion that having a clear hostile interpretation bias may not be a necessary condition to profit from the training on a behavioural level. This implies that a potential selection bias - in the sense that the present sample may have consisted of too few participants with actual aggression problems or hostile interpretation biases - would not have mattered greatly in

outcome. Furthermore, this all does not alter the fact that no effects were found in the present study, even after adjusting for baseline levels of aggression and performance on the emotion perception task.

A third limitation in the present study was that the same faces were applied in both the training procedure and follow-up assessment ("training to the test"), so there is no way of knowing if any generalization to other (real-life) faces and social encounters outside the training context took place. In an earlier study in a student sample with the same facial stimuli as the present study, generalization of the learned target emotion did take place (Griffiths, Jarrold, Penton-Voak, & Munafò, 2015). In another recent study, also conducted in a normal population, was also found that generalization to other faces does take place, as long as this is within the same target emotion (Dalili, Schofield-Toloza, Munafò, & Penton-Voak, 2017). This indicates that it's likely that generalization should take place when using these stimuli in healthy individuals, but it's not certain that this process passes in an equivalent manner in clinical forensic populations.

A fourth aspect to consider concerns the type of stimuli used in the training. Although faces with ambiguous expressions were applied in the training procedure, which is already rather subtle and intricate, these were still all static pictures. In social encounters in daily life, however, emotional expressions are often even more complex. For example, facial expressions often change rapidly and are presented along with co-occurrent verbal or posture cues, which makes generalization of the training task to daily life even more difficult (Schönenberg et al., 2014). One could therefore argue that the applied training paradigm is too unilateral and fails to do justice to the complex nature of all social cues that need to be processed in interactions.

One final remark is in place with regard to the fact that medication use was not included as a potential confounder in the analyses. It could be argued that, inspite of the

randomization, medication use can still have played a confounding role, for example because sedatives can reduce the effectivity of the training or that this can suppress aggression and therefore mask a true training effect.

Research Implications

Because of the above mentioned critical remarks, the generalization of perception of happiness to faces in daily life is an important factor to consider in future studies. This could be addressed by incorporating new faces into the follow-up assessment. Also, including multiple different, non-static faces into the training procedure itself may enhance the chance of generalization taking place. In addition, actual interventions should incorporate multiple aspects of social information processing, such as was originally described by Crick and Dodge (1994).

Clinical Implications

Even though one should first understand which dysfunctions exist in each separate step of social information processing to be able to develop specifically targeted interventions, it may very well be that the power of such interventions ultimately lies in the combined approach towards multiple targets, at least with adults with such long-lasting and profound problems. One possible way to do so in this regard would be to incorporate training strategies for modification of hostile intepretation biases in a more interactive and lively environment in which other strategies are trained simultaneously as well, which can be realised in a virtual reality environment (for an example of such a protocol, see Danique Smeijers and Koole (2019)). In doing so, it is of great importance to compare training effects of such new intervention strategies to those of traditional cognitive behavioural interventions. Do they actually add anything of substance to the present approaches in aggression treatment? Could

the one replace the other, or can they enhance each other's benefits? Those are questions not yet addressed in the present work.

Conclusions

In contrast to earlier studies with adolescents (Penton-Voak et al., 2013; Stoddard et al., 2016), the training procedure in the present study failed to contribute to a decline in aggression or an increase in prosocial behaviour in male adult offenders. Nevertheless, the fact that there was a strong training effect in the perception of happiness over anger, regardless of intelligence levels, seems promising for the development of future interventions in forensic populations. There are certainly important strengths to be found in the present study, such as the strong double blind experimental design and the clear theoretical basis for the intervention. Some limitations and considerations for future studies have also been addressed, of which the influence of age and generalization of perception of happiness to daily life are the most important.

References

AlMoghrabi, N., Huijding, J., & Franken, I. H. A. (2018). The effects of a novel hostile interpretation bias modification paradigm on hostile interpretations, mood, and aggressive behavior. *J Behav Ther Exp Psychiatry, 58*, 36-42. doi:10.1016/j.jbtep.2017.08.003

Angus, D. J., Schutter, D. J. L. G., Terburg, D., Van Honk, J., & Harmon-Jones, E. (2016). A review of social neuroscience research on anger and aggression. In E. Harmon-Jones, M. Inzlicht, E. Harmon-Jones, & M. Inzlicht (Eds.), *Social neuroscience: Biological approaches to social psychology.* (pp. 223-246). New York, NY, US: Routledge/Taylor & Francis Group.

Bousardt, A. M. C., Hoogendoorn, A. W., Noorthoorn, E. O., Hummelen, J. W., & Nijman, H. L. I. (2016). Predicting inpatient aggression by self-reported impulsivity in forensic psychiatric patients. *Criminal Behaviour and Mental Health, 26*(3), 161-173. doi:10.1002/cbm.1955

Cooney, P., Tunney, C., & O'Reilly, G. (2017). A systematic review of the evidence regarding cognitive therapy skills that assist cognitive behavioural therapy in adults who have an intellectual disability. *Journal of Applied Research in Intellectual Disabilities*. doi:10.1111/jar.12365

Crick, N. R., & Dodge, K. A. (1994). A review and reformulation of social information-processing mechanisms in children's social adjustment. *Psychological Bulletin, 115*(1), 74-101. doi:10.1037/0033-2909.115.1.74

Dalili, M. N., Schofield-Toloza, L., Munafò, M. R., & Penton-Voak, I. S. (2017). Emotion recognition training using composite faces generalises across identities but not all emotions. *Cognition and Emotion, 31*(5), 858-867. doi:10.1080/02699931.2016.1169999

Dean, G. (2014). *Neurocognitive risk assessment for the early detection of violent extremists*. New York, NY, US: Springer Science + Business Media.

DeWall, C. N., & Anderson, C. A. (2011). The general aggression model. In P. R. Shaver & M. Mikulincer (Eds.), *Human aggression and violence: Causes, manifestations, and consequences.* (pp. 15-33). Washington, DC, US: American Psychological Association.

Fazel, S., Xenitidis, K., & Powell, J. (2008). The prevalence of intellectual disabilities among 12000 prisoners--A systematic review. *International Journal of Law and Psychiatry, 31*(4), 369-373. doi:10.1016/j.ijlp.2008.06.001

Griffiths, S., Jarrold, C., Penton-Voak, I. S., & Munafò, M. R. (2015). Feedback training induces a bias for detecting happiness or fear in facial expressions that generalises to a novel task. *Psychiatry Res, 230*(3), 951-957. doi:10.1016/j.psychres.2015.11.007

Hornsveld, R. H. J., Muris, P., & Kraaimaat, F. W. (2011). The Novaco Anger Scale–Provocation Inventory (1994 version) in Dutch forensic psychiatric patients. *Psychological Assessment, 23*(4), 937-944.

Hornsveld, R. H. J., Nijman, H. L. I., Hollin, C. R., & Kraaimaat, F. W. (2007). Development of the Observation Scale for Aggressive Behavior (OSAB) for Dutch forensic psychiatric inpatients with an antisocial personality. *International Journal of Law and Psychiatry, 30*(6), 480-491. doi:10.1016/j.ijlp.2007.09.009

Kaal, H. L., Nijman, H. L. I., & Moonen, X. M. H. (2015). Identifying offenders with an intellectual disability in detention in The Netherlands. *Journal of Intellectual Disabilities and Offending Behaviour, 6*(2), 94-101. doi:doi:10.1108/JIDOB-04-2015-0008

Kaal, H. L., Nijman, H.L.I., & Moonen, X.M.H. (2015). *SCIL. Voor volwassenen (SCIL18+) en jongeren van 14 tot en met 17 jaar (SCIL 14-17). Handleiding*. Amsterdam: Hogrefe.

Kobes, M. H. B. M., Nijman, H. H. L. I., & Bulten, E. B. H. (2012). Assessing aggressive behavior in forensic psychiatric patients: Validity and clinical utility of combining two instruments. *Archives of Psychiatric Nursing, 26*(6), 487-494. doi:10.1016/j.apnu.2012.04.004

Kuin, N. C., Masthoff, E. D. M., Munafò, M. R., & Penton-Voak, I. S. (2017). Perceiving the evil eye: Investigating hostile interpretation of ambiguous facial emotional expression in violent and non-violent offenders. *PLoS ONE, 12*(11), e0187080. doi:10.1371/journal.pone.0187080

Lee, B. X. (2016). Causes and cures IX: Consequences of violence. *Aggression and Violent Behavior, 30*, 110-114. doi:10.1016/j.avb.2016.06.013

Lee, B. X., Kjaerulf, F., Turner, S., Cohen, L., Donnelly, P. D., Muggah, R., . . . Gilligan, J. (2016). Transforming Our World: Implementing the 2030 Agenda Through Sustainable Development Goal Indicators. *J Public Health Policy, 37*(Suppl 1), 13-31. doi:10.1057/s41271-016-0002-7

Lisle, A. M. (2007). Assessing learning styles of adults with intellectual difficulties. *Journal of Intellectual Disabilities, 11*(1), 23-45. doi:10.1177/1744629507073997

Marotta, P. L. (2017). A systematic review of behavioral health interventions for sex offenders with intellectual disabilities. *Sexual Abuse: Journal of Research and Treatment, 29*(2), 148-185.

Marsh, A. A., & Blair, R. J. R. (2008). Deficits in facial affect recognition among antisocial populations: A meta-analysis. *Neuroscience and Biobehavioral Reviews, 32*(3), 454-465.

McGillivray, J. A., Gaskin, C. J., Newton, D. C., & Richardson, B. A. (2016). Substance Use, Offending, and Participation in Alcohol and Drug Treatment Programmes: A Comparison of Prisoners with and without Intellectual Disabilities. *Journal of Applied Research in Intellectual Disabilities, 29*(3), 289-294. doi:10.1111/jar.12175

McLeod, A. I. (1985). Remark AS R58: a remark on algorithm AS 183. An efficient and portable pseudo-random number generator. *Journal of the Royal Statistical Society. Series C (Applied Statistics), 34*(2), 198-200.

McNair, L., Woodrow, C., & Hare, D. (2017). Dialectical Behaviour Therapy [DBT] with People with Intellectual Disabilities: A Systematic Review and Narrative Analysis. *Journal of Applied Research in Intellectual Disabilities, 30*(5), 787-804. doi:10.1111/jar.12277

Mikton, C. R., Butchart, A., Dahlberg, L. L., & Krug, E. G. (2016). Global Status Report on Violence Prevention 2014. *American Journal of Preventive Medicine, 50*(5), 652-659. doi:10.1016/j.amepre.2015.10.007

Nijman, H. L. I., Kaal, H. L., Scheppingen, L. v., & Moonen, X. M. H. (2018). Development and testing of a Screener for Intelligence and Learning Disabilities (SCIL). *Journal of Applied Research in Intellectual Disabilities, 31*(1), e59-e67. doi:10.1111/jar.12310

Novaco, R. W. (1994). Anger as a risk factor for violence among the mentally disordered. In J. Monahan & H. J. Steadman (Eds.), *Violence and mental disorder: Developments in risk assessment.* (pp. 21-59). Chicago, IL, US: University of Chicago Press.

Penton-Voak, I. S., Thomas, J., Gage, S. H., McMurran, M., McDonald, S., & Munafò, M. R. (2013). Increasing recognition of happiness in ambiguous facial expressions reduces anger and aggressive behavior. *Psychological Science, 24*(5), 688-697.

Ricciardelli, R., & Memarpour, P. (2016). 'I was trying to make my stay there more positive': rituals and routines in Canadian prisons. *Criminal Justice Studies, 29*(3), 179. doi:10.1080/1478601X.2016.1189423

Schönenberg, Christian, S., Gaußer, A. K., Mayer, S. V., Hautzinger, M., & Jusyte, A. (2014). Addressing perceptual insensitivity to facial affect in violent offenders: First evidence for the efficacy of a novel implicit training approach. *Psychological Medicine, 44*(5), 1043-1052.

Schönenberg, & Jusyte, A. (2014). Investigation of the hostile attribution bias toward ambiguous facial cues in antisocial violent offenders. *European Archives of Psychiatry and Clinical Neuroscience, 264*(1), 61-69. doi:10.1007/s00406-013-0440-1

Schulz, K. F., Altman, D. G., & Moher, D. (2010). CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. *Bmj, 340*, c332. doi:10.1136/bmj.c332

Schwenk, C., Gensthaler, A., Romanos, M., Freitag, C., Schneider, W., & Taurines. (2014). Emotion recognition in girls with conduct problems. *European Child and Adolescent Psychiatry, 23*, 13-22.

Smeijers, D., & Koole, S. L. (2019). Testing the Effects of a Virtual Reality Game for Aggressive Impulse Management (VR-GAIME): Study Protocol. *FRONTIERS IN PSYCHIATRY, 10*, 83.

Smeijers, D., Rinck, M., Bulten, E., Van den Heuvel, T., & Verkes, R. J. (2017). Generalized hostile interpretation bias regarding facial expressions: Characteristic of pathological aggressive behavior. *Aggressive Behavior, 43*(4), 386-397.

Stoddard, J., Sharif-Askary, B., Harkins, E. A., Frank, H. R., Brotman, M. A., Penton-Voak, I. S., . . . Leibenluft, E. (2016). An open pilot study of training hostile interpretation bias to treat disruptive mood dysregulation disorder. *Journal of Child and Adolescent Psychopharmacology, 26*(1), 49-57. doi:10.1089/cap.2015.0100

Verhage, F. (1964). *Intelligentie en leeftijd. Onderzoek bij Nederlanders van twaalf tot zevenenzeventig jaar*. Assen, Netherlands: Van Gorcum.

Wichmann, B. A., & Hill, I. D. (1982). Algorithm AS 183: An efficient and portable pseudo-random number generator. *Journal of the Royal Statistical Society. Series C (Applied Statistics), 31*(2), 188-190.

Wistedt, B., Rasmussen, A., Pedersen, L., Malm, U., Traskman-Bendz, L., Wakelin, J., & Bech, P. (1990). The development of an observer-scale for measuring social dysfunction and aggression. *Pharmacopsychiatry, 23*(6), 249-252. doi:10.1055/s-2007-1014514

Table 1
Descriptive data, including baseline measures, of the participants in the two study groups who were included in statistical analysis

	Intervention Baseline (*n* = 45)	Control Baseline (*n* = 41)	Intervention After training (*n* = 45)	Control After training (*n* = 41)	Intervention Follow-up (*n* = 39)	Control Follow-up (*n* = 36)
Age (mean, *sd*)	37.3 (11.0)	41.9 (11.6)				
Educational level[1] (median, range)	3 (1-5)	4 (1-5)				
Estimated IQ <85 based on SCIL[2] (*n*, % of group)	22 (49%)	15 (37%)				
Currently detained for a violent crime (*n*, % of group)	35 (78%)	28 (68%)				
Total number of convictions for non-violent crimes (mean, *sd)*	18.8 (26.3)	15.0 (22.7)				
Number of convictions for violent crimes (mean, *sd*)	7.93 (9.1)	8.2 (10.1)				
Novaco Anger Scale (mean, *sd*)	83.0 (15.0)	76.9 (17.6)	79.33 (14.4)	76.6 (16.8)	78.0 (11.8)	71.8 (13.6)
Provocation Inventory (mean, *sd*)	48.1 (11.2)	44.8 (13.4)	46.6 (11.4)	44.3 (11.6)	45.9 (9.1)	42.7 (11.6)
Weekly Self-Report[3] (mean, *sd*)	6.5 (5.2)	6.1 (5.4)	4.7 (4.0)	6.0 (8.7)		
OSAB[3,4] Irritation/Anger (mean, *sd*)	8.6 (2.2)	8.2 (2.7)	7.8 (2.1)	8.7 (4.7)		
OSAB Aggressive Behavior (mean, *sd*)	13.1 (3.5)	12.2 (3.3)	12.7 (3.4)	11.9 (3.2)		

OSAB Prosocial Behavior (mean, *sd*)	32.3 (5.2)	31.0 (8.3)	30.6 (5.6)	29.8 (8.1)		

Note. [1] educational level was based on the classification system of Verhage (1964) in Dutch education with 6 levels of education: (1) not graduated from primary school, (2) only graduated from primary school, (3) vocational education, (4) secondary vocational education, (5) higher vocational education, (6) academic education.
[2] SCIL = Screener for Intelligence and Intellectual Disability
[3] Both for OSAB and self-report the after training mean score reflects a mean of weekly measures during the six weeks after training
[4] OSAB = Observation Scale for Aggressive Behavior

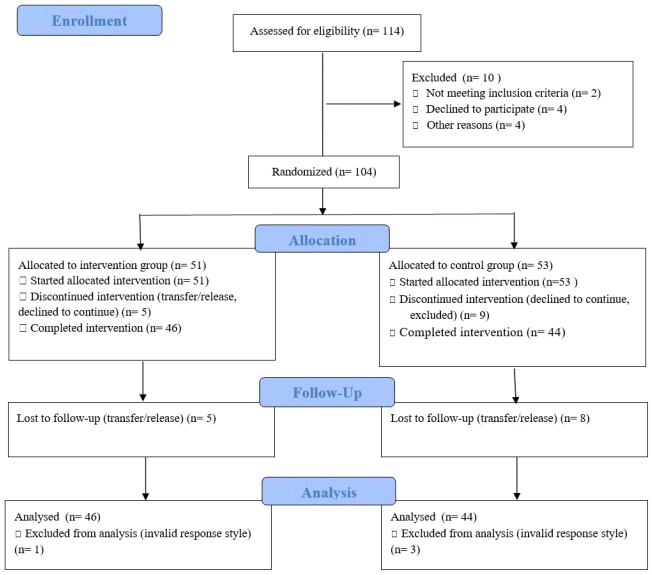


Figure 1. Flowchart of the inclusion process for the Experimental Group (EG) and Control Group (CG) according to CONSORT 2010 guidelines (Schulz, Altman, & Moher, 2010). Also displayed are drop-out rates, mostly related to unforeseen transfer or release, as well as to participants' refusal to continue. Reasons for drop-out during the training week include illness, too much disruptive behaviour, or correctional measures for rule breaking during the training week.

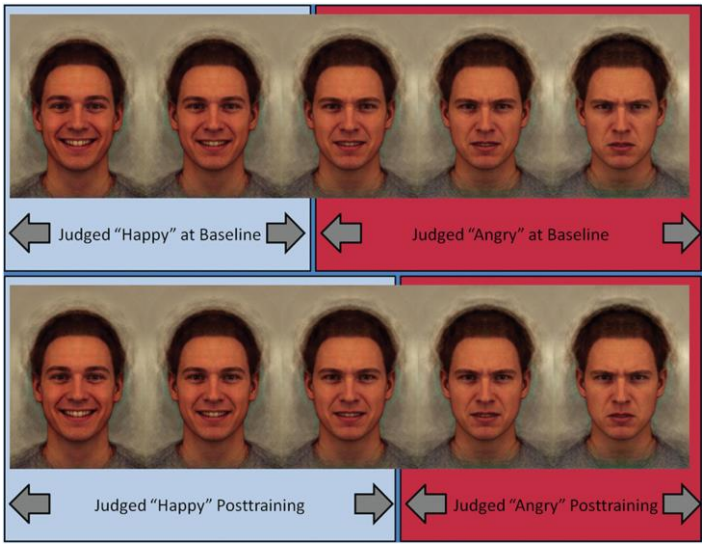
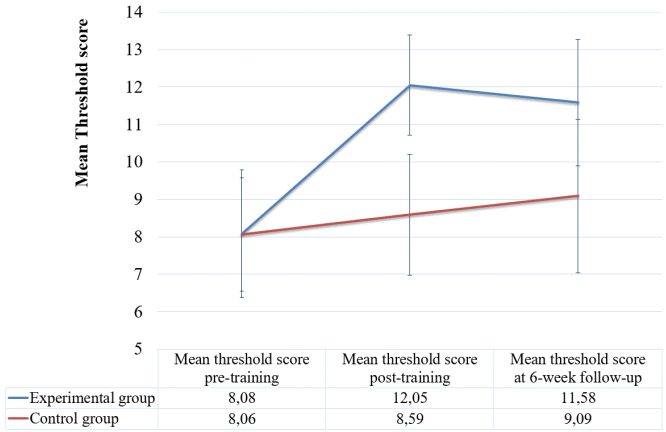


Figure 2. Illustration of the stimuli and design of the intervention, portraying how the balance point (threshold) between happy and angry responses may shift towards a larger proportion of happy responses after treatment in the intervention group (bottom panel), compared to the baseline balance point (top panel) in this group.



	Mean threshold score pre-training	Mean threshold score post-training	Mean threshold score at 6-week follow-up
Experimental group	8,08	12,05	11,58
Control group	8,06	8,59	9,09

Figure 3. Mean threshold scores on the emotion perception task with error bars representing standard deviations for the intervention and control groups at baseline, directly after the training and at six weeks follow-up.

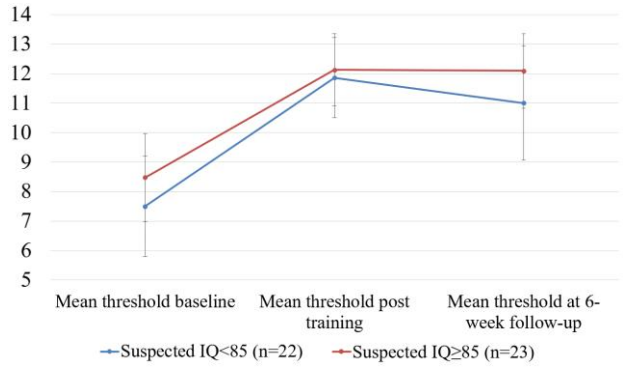


Figure 4. Mean threshold scores and error bars representing standard deviations at baseline, directly post training and at 6-week follow-up of intervention-group participants with versus without estimated mild intellectual disability (based on scores on the Screener for Intelligence and Intellectual Disability).

Table 2
Minimally and Fully adjusted linear regression coefficients of LOG-transformed values of post-training measures of aggression, hostility and prosocial behaviour

	Minimally adjusted [a]		Fully adjusted [b]	
	B [95% CI]	*p*-value	*B* [95% CI]	*p*-value
Novaco Anger Scale [1]	0.01 [-0.01, 0.03]	0.40	0.01 [-0.01, 0.03]	0.39
Novaco Anger Scale 6w. [1]	-0.02 [-0.05, 0.01]	0.18	-0.02 [-0.05, 0.01]	0.13
Provocation Inventory [1]	0.004 [-0.03, 0.04]	0.81	0.002 [-0.03, 0.04]	0.89
Provocation Inventory 6w. [1]	-0.01 [-0.04, 0.02]	0.40	-0.02 [-0.05, 0.01]	0.30
Weekly Self-Report [2]	0.06 [-0.04, 0.15]	0.25	0.05 [-0.05, 0.15]	0.29
OSAB Irritation/Anger [2,3]	0.04 [-0.01, 0.09]	0.11	0.03 [-0.02, 0.08]	0.20
OSAB Aggressive Behavior	-0.003 [-0.04, 0.03]	0.84	-0.01 [-0.04, 0.03]	0.60
OSAB Prosocial Behavior	-0.01 [-0.05, 0.03]	0.63	-0.01 [-0.05, 0.03]	0.56

Note. [a] Outcomes of the linear regression analyses, minimally adjusted for baseline
[b] Outcomes of the linear regression analyses, fully adjusted for baseline, age, intelligence (Screener for Intelligence and Intellectual Disability) and age x condition
[1] The Novaco Anger Scale and Provocation Inventory were assessed directly after training (first score) and after six weeks (second score) and compared to the baseline assessment
[2] Weekly self-report and OSAB scores were based on average scores of the four ratings prior to training and average scores of ratings during the six weeks after training.
[3] OSAB = Observation Scale for Aggressive Behavior